Reducing hallucination in structured outputs via Retrieval-Augmented Generation

Year: 2024 | Citations: 108 | Authors: Patrice Bechard, Orlando Marquez Ayala

---

Abstract

A common and fundamental limitation of Generative AI (GenAI) is its propensity to hallucinate. While large language models (LLM) have taken the world by storm, without eliminating or at least reducing hallucinations, real-world GenAI systems may face challenges in user adoption. In the process of deploying an enterprise application that produces workflows based on natural language requirements, we devised a system leveraging Retrieval Augmented Generation (RAG) to greatly improve the quality of the structured output that represents such workflows. Thanks to our implementation of RAG, our proposed system significantly reduces hallucinations in the output and improves the generalization of our LLM in out-of-domain settings. In addition, we show that using a small, well-trained retriever encoder can reduce the size of the accompanying LLM, thereby making deployments of LLM-based systems less resource-intensive.